STAT 621 Chapter 4

Zepu Zhang
October 17, 2011

In a contingency table, observations are counts and they are organized in a table.

In the McNemar test for significance of changes (Chapter 3), observations are also counts organized in tables. The observations are made on one single sample twice, under two different conditions, e.g. before and after a treatment.

In many other applications, it is impossible or meaningless (for the purpose of the study) to make observations twice on a single sample. Instead, observations are made on samples drawn from two (or more) populations. This situation is addressed in this chapter.

1 2×2 contingency tables

Take two samples from two populations.

Assumptions:

1. Each sample is a random sample.
2. The two samples are mutually independent.
3. Each observation may be classified into either "class 1" or "class 2".

	class 1	class 2	total
population 1	O_{11}	O_{12}	n_1
population 2	O_{21}	O_{22}	n_2
total	c_1	c_2	$N = n_1 + n_2 = c_1 + c_2$

There are 3 situations for a 2×2 contingency table:

1. Row totals are fixed (not random), column totals are random outcomes. Use the χ^2 test.
2. Both row and column totals are fixed. Use Fisher's Exact Test.

3. Both row and column totals are random. This situation is the same as the more general $r \times c$ contingency tables, and will be discussed in the next section.

   When the row totals or column totals are random, more appropriate symbols appear to be the upper-case N_i and C_i. We do not make such strict distinctions in the notation since we discuss both random and fixed cases.

1.1 The χ^2 test for differences in probabilities

Let

$$p_i = P(\text{an observation in population i belongs to class 1}), \quad i = 1, 2$$

The goal is to test whether $p_1 = p_2$. Hence

$$H_0: \ p_1 = p_2$$

H_a may be two-sided or one-sided.

Let's estimate p_1 by O_{11}/n_1 and p_2 by O_{21}/n_2. To compare p_1 and p_2, we examine $O_{11}/n_1 - O_{21}/n_2$. If H_0 is true, this difference should be fluctuating around 0. We need to know its null distribution in order to judge whether the observed value is significantly nonzero.

The exact distribution is difficult to get, so let's use large sample approximations. With large samples,

$$\frac{O_{11}}{n_1} - \frac{O_{21}}{n_2} \sim N\left(p_1 - p_2, \ \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$$

Under H_0, we use the estimator $p_1 = p_2 = c_1/N$. Then

$$\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2} = \frac{c_1}{N}\left(1 - \frac{c_1}{N}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right) = \frac{c_1 c_2}{n_1 n_2 N}$$

Subsequently, standardize to get a standard normal test statistic:

$$T = \frac{\frac{O_{11}}{n_1} - \frac{O_{21}}{n_2}}{\sqrt{\frac{c_1 c_2}{n_1 n_2 N}}} = \frac{\sqrt{N}\left(n_2 O_{11} - n_1 O_{21}\right)}{\sqrt{n_1 n_2 c_1 c_2}} = \frac{\sqrt{N}\left(O_{11}O_{22} - O_{12}O_{21}\right)}{\sqrt{n_1 n_2 c_1 c_2}}$$

If H_a is two-sided, we can carry out the test by checking whether T^2 is too big. The null distribution of T^2 is χ^2_1.

Example EX 1, page 182.

Example EX 2, page 183.

1.2 Fisher's exact test

Example EX 3, page 190.

In this case, both row totals and column totals are pre-determined: they are not random.

Let's examine the null distribution of O_{11}. Under H_0, each of the N subjects has the same chance of being in class 1 (column 1). The "experiment" is that we randomly pick c_1 out of the (perfectly mixed) N subjects, and O_{11} of those picked are from population 1.

This is the same experiment as the following: given N balls in which n_1 are red and $n_2 = N - n_1$ are blue; randomly pick c_1 balls. What is the distribution of the number of red balls picked? The distribution of the number, ranging between 0 and $\min(n_1, c_1)$, is called "hypergeometric".

The probability of O_{11} is given by

$$P(O_{11}) = \frac{\binom{n_1}{O_{11}}\binom{n_2}{O_{21}}}{\binom{N}{c_1}}, \quad O_{11} = 0, \ldots, \min(n_1, c_1)$$

Because both row totals and column totals are fixed, the value of O_{11} determines the counts in all the other cells. Hence we can write

$$P(O_{11}) = \frac{\binom{n_1}{O_{11}}\binom{n_2}{c_1 - O_{11}}}{\binom{N}{c_1}}, \quad O_{11} = 0, \ldots, \min(n_1, c_1)$$

To carry out the test, we calculate the p-value by adding the probabilities of O_{11} taking the observed and more extreme values. (Also, of course, note whether the H_a is two-sided or one-sided.)

Large sample approximation: use the known results about the mean and variance of the hypergeometric distribution to standardize O_{11} and get a standard normal test statistic.

Example EX 3, page 190.

2 $r \times c$ contingency tables

Two-way contingency table:

	class 1	class 2	$\cdots$	class c	total
population 1	O_{11}	O_{12}	$\cdots$	O_{1c}	n_1
population 2	O_{21}	O_{22}	$\cdots$	O_{2c}	n_2
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
population r	O_{r1}	O_{r2}	$\cdots$	O_{rc}	n_r
total	c_1	c_2	$\cdots$	c_c	$N = \sum n_i = \sum c_i$

> When the row or column totals are random, more appropriate symbols appear to be the upper-case N_i, or R_i, and C_i. Since we discuss both random and non-random situations, we do not change notation according to this aspect.

With pleasant consistency, we will use "Pearson's chi-squared statistic"

$$T = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \quad \text{where } E_{ij} = n_i \frac{c_j}{N} = N \frac{n_i}{N} \frac{c_j}{N}.$$

Here E_{ij} is the "expected value" of the variable in the cell (i, j) under certain null hypothesis. We see the null hypothesis is that the class fractions are common across populations, or some sort of independence between the row and column effects.

All tests in this section are two-sided, hence we'll safely use a χ^2 test statistic. The two-sidedness is because $r > 2$ or $c > 2$, or both $r > 2$ and $c > 2$; a one-sided hypothesis is meaningless or of little interest.

2.1 The χ^2 test for differences in probabilities

One sample is drawn independently from each population. Each object is classified as one of the c classes.

Row totals are fixed but column totals are random.

Let p_{ij} be the probability that a randomly selected object in the ith population belongs to the jth class. The goal is to test whether objects are allocated to the classes by the same relative proportions in all the populations, i.e. whether the discrete distribution (c possible values) is the same for all populations. Hence

$$H_0 : \; p_{1j} = p_{2j} = \cdots = p_{rj} \quad \text{for } j = 1, \ldots, c$$

In the test statistic, we see that E_{ij} is n_i times the p_{ij} estimated by $\frac{c_j}{N}$ (the same for all i).

Approximate null distribution in large-sample situations:

$$T \sim \chi^2_{(r-1)(c-1)}$$

Example EX 1, page 202.

Note When is the χ^2 approx to the true null distribution satisfactory? General answer: if the E_{ij}s are not too small. Empirical rule of thumb: all E_{ij}s are > 0.5 and at least half are > 1.0. What if this requirement is not met? Combine some classes (if meaningful).

Exercise Verify that the χ^2 test statistic obtained for the 2×2 table, $\frac{N(O_{11}O_{22}-O_{12}O_{21})^2}{n_1 n_2 c_1 c_2}$, is the T above. (First notice the degree of freedom is correct: $(2-1)(2-1) = 1$.) (I have not checked this yet. Maybe the "convenient form" (5), page 200, is useful.)

Question It may not be obvious that the d.f. is $(r-1)(c-1)$. But why is it not rc? (This is obvious.)

2.2 The χ^2 test for independence

Suppose a random sample of size N is obtained. Each object is classified according to two criteria. By criterion 1, it belongs to one of r classes, whereas by criterion 2 it belongs to one of c classes. The count of objects in the ith class by criterion 1 and the jth class by criterion 2 is O_{ij}.

Both row and column totals are random.

The goal is to test whether the two classification criteria are "independent". That is, taking a random object, which row it belongs and which column it belongs do not affect each other. In statistical terms,

$$H_0 : \; P(\text{row } i, \text{ column } j) = P(\text{row } i) \cdot P(\text{column } j)$$

In this case, it is immediately understandable that E_{ij} is estimated (or defined) by $N(n_i/N)(c_j/N)$.

Approximate null distribution:

$$T \sim \chi^2_{(r-1)(c-1)}$$

Example EX 2, page 206.

2.3 The χ^2 test with fixed marginal totals

Both row and column totals are fixed, i.e. pre-determined.

Hypotheses: they are essentially the same as the last two tests, in terms of "probabilities" or "independence". The actual wording is often tailored to specified problems.

Approximate null distribution:

$$T \sim \chi^2_{(r-1)(c-1)}$$

The exact null distribution can be derived in a way similar to Fisher's exact test.

Example EX 3, page 210.

Example EX 4, page 211.

2.4 The median test

The median test is designed to examine whether several samples come from populations having the same median. This is a special form of the χ^2 test with fixed row and column totals.

Make a $r \times 2$ table. O_{i1}, $i = 1, \ldots, r$, is the number of observations in the ith sample are are below the grand median, whereas O_{i2} is the number above.

Because of the way the grand median is used, $c_1 \approx c_2 \approx N/2$. Hence both the row and column totals are fixed.

The test is whether $p_{11} = \cdots = p_{r1}(= 0.5)$. Or equivalently,

$$H_0 : \text{ all } r \text{ populations have the same median}$$

Using the same test statistic, the approximate null distribution is

$$T \sim \chi^2_{r-1}$$

Example EX 1, page 220.

Exercise Compare with the two-sample sign test.

The same idea can be used to test whether the populations have the same, say, 25th percentile. Or indeed, we can simultaneously test multiple quantiles, which constitutes a test that the populations have the same (rough) distribution.

3 Measures of dependence

The tests introduced in the last section concern whether a distribution changes with population or whether two classification criteria interact with each other. Either way, the test is about whether the rows and columns are "independent". Instead of tests, one may want a measure for the strength of the dependency.

3.1 Cramer's contingency coefficient

Take the χ^2 test statistic T and standardize it by its max possible value so that the result has a known, finite range, $[0, 1]$ in particular.

T achieves its max when each row and each column has at most one non-zero cell. The max value is $N(\min(r, c) - 1)$. Define

$$\text{Cramer's coef} = \sqrt{\frac{T}{N(\min(r, c) - 1)}}$$

This measure is "scale invariant", that is, if all cell values are, say, 10 times larger, the measure does not change.

3.2 The phi coefficient

For a 2×2 table, the Cramer's coef turns out to be

$$\sqrt{\frac{(O_{11}O_{22} - O_{12}O_{21})^2}{n_1 n_2 c_1 c_2}}$$

In this situation, it makes sense to talk about the "nature" (or direction) of the dependence (or "association"): are rows and columns positively or negatively associated?

To reflect this nature, one wants to keep the sign of $O_{11}O_{22} - O_{12}O_{21}$, hence the phi coef is defined as

$$\frac{O_{11}O_{22} - O_{12}O_{21}}{\sqrt{n_1 n_2 c_1 c_2}}$$

Example EX 7, page 236.

4 χ^2 goodness-of-fit test

In the previous tests, the hypotheses are about particular aspects of the distribution of the sample, for example the median, the probability of one or a few classes.

In this section, we test whether the data sample "fits" a specific (entire) distribution, e.g. normal, or exponential.

We split the sample into c classes; let the count of the sample in class i be O_i. Then we compute the expected value under the hypothesized distribution, E_i, and form the test statistic

$$T = \sum_{i=1}^{c} \frac{(O_i - E_i)^2}{E_i}$$

Approximate null distribution is

$$T \sim \chi^2_{c-1}$$

If the hypothesized distribution is discrete, the "class" terminology above is natural. If it is continuous, we divide the value range into "intervals" and the test works the same way.

Example EX 1, page 242.

Note If the hypothesized distribution needs k parameters (for example normal distribution needs mean and variance for the distribution to be fixed), we first estimate the parameters <u>using a good (traditional, standard) estimation method</u> from the data, then use the approximate null distribution $T \sim \chi^2_{c-1-k}$.

Example EX 2, page 244.

Example EX 3, page 246.

Note This is the same test as the χ^2 test for differences in probabilities based on a $2 \times c$ table. Imagine we construct a test this way: row 1 is based on the data sample; row 2 is based on a huge sample from the hypothesized distribution. Because the sample size is huge, we have c_j/N equal to the theoretical value and $O_{2j} - E_{2j} = 0$. Also verify that the degree of freedom of the χ^2 is $(c-1)(2-1) = c-1$.

Note Recall our comments on the median test. The median test can be extended to test that the two populations have the same set of quantiles. That is the same idea, and it is more general in that it tests whether two samples come from populations with the same distribution (hence the hypothesized distribution does not need to be a standard one).